How Many Dimensions Do You Actually Need? Testing Embedding-Dimension Rules on Their Slope, Not Their Value

Anonymous authors
Paper under double-blind review

Abstract

Rules of the form $n > c \ln N$, where N counts the items to be represented, are widely used to choose an embedding dimension. They are usually validated by evaluating the rule at one N and comparing it with a conventional width, a check that a wide family of constants and functional forms passes equally well. In this work, we measure how the required dimension actually moves with N, holding the task fixed and varying N over up to 2.7 decades in associative recall with an attention head, identity retrieval and topic recovery in word embeddings, and node classification on citation graphs. We find that the dimension a task needs is set by what it must distinguish rather than by N. On the same synthetic embeddings with few planted topics, recovering the topics needs as many dimensions as there are topics whatever N is, while telling the words apart needs a width that grows with $\ln N$. We also find that identity tasks on different data grow at rates that differ by more than an order of magnitude, and that classification tasks show no detectable common growth. Along the way, we show that dimension floors estimated relative to a noisy curve's maximum acquire a spurious slope in $\ln N$, and give an estimator that does not. We hope the protocol lets future dimension rules be stated for a task and tested on their slope.

1 Introduction

How wide should an embedding be? The practical answer is a sweep, and the practical finding is that accuracy saturates: past some width, more dimensions neither help nor hurt much. This shape invites a theory of the saturation point, and several have been proposed. The Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984; Dasgupta & Gupta, 2003) gives $O(\ln N)$ for preserving pairwise distances. Yin & Shen (2018) give a bias–variance account through the PIP loss. An entropy argument (Su, 2020) gives the concrete $n > 8.33 \ln N$, which has been adapted to choose the width of graph neural networks (Luo et al., 2021; Yang et al., 2023).

These rules are checked in a characteristic way. One picks a dataset, computes the formula, and observes that the result is close to a dimension people already use: 96 against 100 for word vectors, 52 against 64 for an attention head. The agreement is real but it is weak evidence, because the check has one degree of freedom and the prediction has one. Any quantity of the right order passes. How the requirement moves with N has been measured for particular tasks: dense retrieval needs more width as the index grows (Reimers & Gurevych, 2021; Weller et al., 2026), and graph embeddings show no clear dependence on N (Gu et al., 2021). To our knowledge, no named rule has been tested on its slope with the task held fixed.

The informative question is how the requirement *moves*. A rule $n > c \ln N$ asserts that doubling N adds $c \ln 2 \approx 0.69c$ dimensions, and testing that requires varying N while holding everything else fixed — the task, the supervision, the architecture. In this work, we measure it in four settings, and find that the answer is organised by the task rather than by N (Figure 1). Tasks that must tell the N items apart need a width that grows with N, at rates that differ by more than an order of magnitude from one dataset to another. Tasks that only need to separate K categories behave differently: on synthetic data with few categories they need K dimensions whatever N is, and on real graphs the width shows no common trend with N. The cleanest case is a single set of synthetic embeddings

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

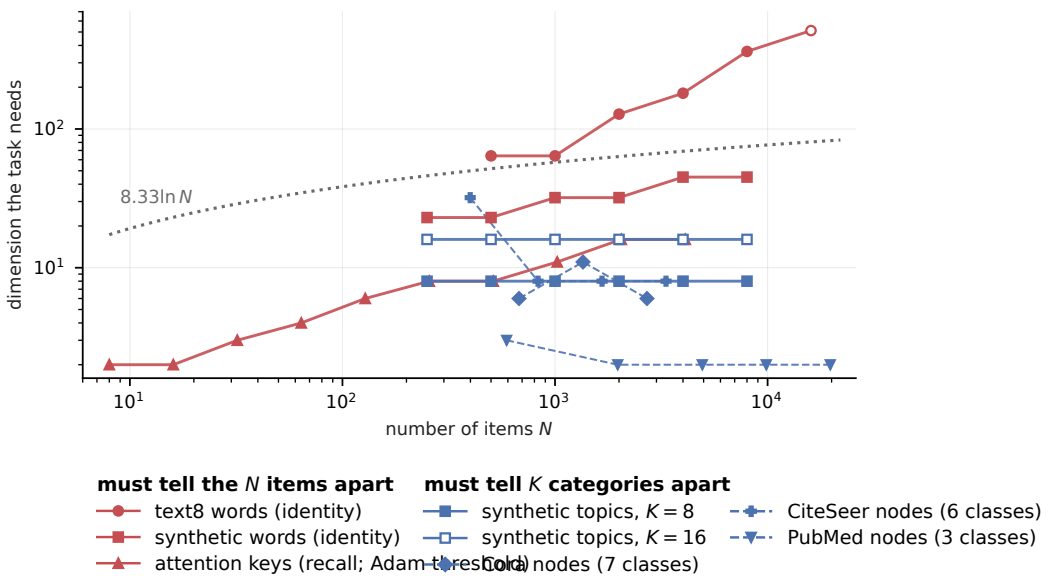

Figure 1: The dimension each task needs against the number of items N, with $8.33 \ln N$ dotted for reference. Red: tasks that must tell the N items apart; blue: tasks that must separate K categories. Each point is the smallest dimension reaching 95% of the task's reference accuracy (§2.1); the hollow circle is censored at the top of the `text8` grid, and PubMed's floors of 2 sit at the bottom of its grid. The red synthetic series and the blue $K = 8$ series are the same embeddings scored on two tasks; at $K = 32$ the two tasks do not separate (Table 1).

scored two ways. With eight or sixteen planted topics, recovering them needs K dimensions at every N, while telling its words apart needs a width that grows with $\ln N$, at a rate compatible with the rule's constant when the words fall into few topics and slower when they fall into many.

Contributions. First, a protocol: we formalise the slope test for rules of the form $n > f(N)$, show why a value check at a single N cannot discriminate between candidate rules, and give an estimator of the requirement whose expectation is zero when the truth is flat (§2.1). Second, a finding: across four settings spanning up to 2.7 decades of N, with the dimension grid extended far below any published floor, the requirement tracks what the task must distinguish. Topic recovery needs K dimensions at every N when K is small, the two real identity tasks grow with N at rates that differ by a factor of 42, and node classification shows no common trend across graphs. No single constant c describes them (§3, Tables 1–3). Third, a practical pitfall and its remedy. A requirement defined relative to a noisy curve's own maximum carries a bias that shrinks with N, and so produces a spurious negative slope. We measure the bias and give the fixed-reference estimator that removes it (§3.4, Figure 4).

What this paper does not claim. This work does not aim to show that any particular dimension rule or selector is wrong, nor does it question the entropy argument, whose rate is compatible with the identity task it describes on clean data with few topics (§3.1). Rather, it is intended to provide a protocol under which such rules can be stated for a task and compared on how the requirement scales.

2 The Rule, the Slope Test, and the Settings

The entropy argument. Su (2020) models a skip-gram embedding as $p(w_i, w_j) \propto \exp\langle u_i, v_j \rangle$ over N^2 pairs. Assuming each coordinate has magnitude ≈ 1, so $\|u\| \approx \sqrt{n}$, and that vectors are uniform on the sphere, the model entropy is

$$H(n) \;=\; \ln N^2 + h_n, \qquad h_n \;\to\; -\tfrac{1}{2} \ln \tfrac{\sqrt{5}+1}{2}\, n \;\approx\; -0.2406\, n. \tag{1}$$

We verify (1) numerically: h_n/n converges to -0.2406 from -0.2430 at $n = 32$. The stated criterion is $H(n) \le H_{\text{data}}$; since $H_{\text{data}} > 0$, this is weakened to the sufficient condition $H(n) < 0$, giving

$$n \;>\; \ln N^2/0.24 \;=\; 8.33 \ln N. \tag{2}$$

What the rule says, and how it is used. The derivation keeps all N^2 word pairs distinguishable, so it concerns a task that must tell every item apart, and it yields a *sufficient* width for that task: dropping H_{data} makes the condition stronger than the criterion it stands in for. In use the rule has served two further roles — as a minimum below which an embedding is too small, and as a direct prescription, as in the graph selectors of Luo et al. (2021); Yang et al. (2023), which add a structural-entropy term to $\ln N^2$. We test the rule as published; the exact criterion needs H_{data}, which none of these uses estimates. The measurements below speak to each role: a task that needs more than $8.33 \ln N$ bears on sufficiency, one that needs far less bears on the minimum, and the slope bears on the prescription.

Relation to Johnson–Lindenstrauss. The elementary JL bound (Dasgupta & Gupta, 2003) requires $n \ge 4 \ln N/(\epsilon^2/2 - \epsilon^3/3)$, which never falls below $24 \ln N$, even as $\epsilon \to 1$. Equation (2) is about $3\times$ below the worst-case guarantee: it is an average-case entropy-matching criterion, not JL restated.

2.1 The slope test

Let $n^\star(N)$ be the minimum dimension at which a fixed task reaches its own performance plateau. Read as a sufficient width, a rule $n > f(N)$ claims $n^\star \le f(N)$ at each N; read as a prescription or a floor, its content is the slope $\mathrm{d}n^\star/\mathrm{d}\ln N$.

Why value checks are weak. A check at a single N accepts any constant c with $c \ln N \approx n_{\text{conventional}}$. At $N = 10^5$ this interval is wide: constants from 6.3 to 10.4 all land within $\pm 25\%$ of 96. Nor can it separate functional forms: over one decade of N, $c \ln N$ and $c' N^\alpha$ with small α are numerically indistinguishable. Discriminating needs several N and a grid bracketing the collapse.

What we hold fixed. Varying N while holding "the task" fixed is the design problem. We hold the architecture, the optimiser, the training budget and the nominal number of labelled examples per class constant, and vary N. On every run we record the quantities that inevitably co-vary: graph degree, realised labels per class, evaluation-set size, and the class-distribution distance between validation and test splits. §4 reports where this is imperfect.

Locating $n^\star$ without a biased estimator. The natural quantity, "the smallest d within a tolerance of the curve's maximum", is biased in a way that matters for a slope test. The maximum of 13 noisy means exceeds the true plateau by an amount proportional to the seed noise. Wherever seed noise falls with N — it does in our graph arm, by -0.66 accuracy points per unit of $\ln N$ — a flat truth therefore yields a *negative* slope (§3.4). We use two estimators that avoid it. The primary quantity is the gap at a fixed dimension against a *fixed reference* dimension,

$$g_d(N) \;=\; \mathrm{acc}(d_{\text{ref}}, N) - \mathrm{acc}(d, N), \qquad d_{\text{ref}} = 128, \tag{3}$$

a difference of two means whose expectation is zero under a flat truth. A requirement of $8.33 \ln N$ would exceed every $d \le 32$ in every graph cell we keep, so g_d should be large throughout our range, and it should step up where N crosses $\exp(d/8.33)$ ($N \approx 2{,}200$ for $d = 64$). The secondary quantity, used where a single number per N is needed, is the smallest grid dimension reaching 95% of the accuracy at d_{ref} on graphs, and of the curve's peak for attention and word vectors, where the peak is 1.000 wherever the task is solved — a relative tolerance that does not shrink with N — with a seed bootstrap and a flag at a grid edge. Topic recovery peaks below 1, at $d = K$, so its floors are also checked against $\mathrm{acc}(128)$ (§3.1). Every floor is repeated at 98 and 99% (Appendix B).

Inference. All intervals are t-intervals. Graph cells are nested subsamples of three graphs, so pooled regressions are reported with dataset fixed effects and with cluster-robust (CR1) standard errors on t_{G-1}, $G = 3$; the pooled p-value alone would treat 12 cells as independent. Beside each gap slope we report what the same estimator returns on a simulated flat truth with each cell's noise.

The slope test for a rule $n > f(N)$, in full. (1) State the task: what it must distinguish, its metric, and the reference width d_{ref} that defines its plateau. (2) Choose several values of N over as wide a range as the data allow, holding the architecture, optimiser, budget and per-class supervision fixed, and record every quantity that must co-vary with N. (3) Sweep d on a grid that brackets the collapse at every N, with repeated seeds. (4) Read the requirement against d_{ref}, never against the curve's own maximum. (5) Simulate a flat truth with each cell's noise and report its slope beside the measured one. (6) Regress on $\ln N$ with t-intervals, clustering cells that share a source. (7) Repeat at every tolerance; a claim whose direction changes with the tolerance is not a finding.

2.2 SETTINGS

All runs are CPU-only: 3,841 gradient-trained models and 1,003 word-vector rank slices, all read by one analysis script. Full commands, grids, seeds and run counts are in Appendix A.

Word embeddings. (i) A synthetic generator with K planted latent topics and Zipfian frequencies, 3 seeds. Each rank-n embedding is scored on two tasks: *topic recovery*, the Spearman correlation between embedding similarity and planted topic similarity, and *identity retrieval*, whether each word's reconstructed row is closest to its own among all N. (ii) `text8` (17M tokens), window ± 5, positive PMI, truncated SVD (Levy & Goldberg, 2014) — a close relative of (1), whose skip-gram form factorises shifted PMI, here without its norm assumption — scored on identity retrieval. Vocabulary N from 500 to 16,000; rank up to 512. The factorisation is deterministic, so there is one per N; each is scored on two random evaluation subsamples of 2,000 words (all words when $N \leq 2{,}000$).

Attention. Associative recall: N positions each hold a distinct learned key and a random value from a vocabulary of 16; a query repeats one key and the model must emit that key's value, so the head must tell the N keys apart. One head, $d_{\text{model}} = 64$; the swept width is the query/key projection, 12 values $\{1, 2, 3, 4, 6, 8, 11, 16, 23, 32, 48, 64\}$; $N \in \{8, \ldots, 4096\}$; Adam, 2500 steps, 3 seeds. The value path stays 64-wide, so the swept axis is the QK rank rather than a full head. A head of width 2 with hand-set weights can solve this task at any N, so the sweep measures the width at which Adam *finds* a solution: an optimisation threshold rather than a representational floor (§4).

Node classification. Cora, CiteSeer, PubMed (Yang et al., 2016), with 7, 6 and 3 classes. We keep a fraction of the nodes as an induced subgraph and derive a fresh split on it. The split takes up to **20 labelled nodes per class** (fewer when a class has fewer survivors), a validation set of $\min(500, \max(30, 20\%))$ of the remainder capped at half of it, and the rest as test, shuffling the remainder before the validation/test cut (Appendix C). Supervision is thus constant in nominal count while N moves from 399 to 19,717 on the usable cells. GCN (Kipf & Welling, 2017), dimensions $\{2, 3, 4, 6, 8, 11, 16, 23, 32, 45, 64, 91, 128\}$, 5 seeds. Cells with a test split under 200 nodes are excluded: the smallest Cora fraction keeps 20–26 test nodes, a binomial standard error near 10 points.

3 RESULTS

3.1 ON THE SAME EMBEDDINGS, THE TASK SETS THE REQUIREMENT

The synthetic generator isolates the question. The data, the embedding and N are shared between its two tasks; only what the task must distinguish differs. Topic recovery needs K dimensions at every N, to the resolution of the grid (whose next width down is 6, 11 or 23): the floor equals K in 6/6 cells for $K = 8$ and 6/6 for $K = 16$, at every tolerance up to 99%. At $K = 32$ it equals K in 5/6 cells at 95% and 6/6 at 99%; the closest decision is at $N = 500$, where width 23 misses the bar by 1.3×10^{-4}. Read against $\text{acc}(128)$ instead of the peak, the $K = 8$ and 16 floors equal K in 18/18 and 18/18 cells (six N, three tolerances), while $K = 32$ equals K in 12/18 and reads 23 in the rest. Identity retrieval on the same embeddings grows at $+7.25$ $[+3.98, +10.53]$ per unit of $\ln N$ for $K = 8$, $+5.28$ $[+0.32, +10.23]$ for $K = 16$ and $+1.85$ $[-1.12, +4.83]$ for $K = 32$. The first two intervals contain 8.33; the third excludes it and includes 0: on the task its derivation describes, and on clean data, the rule's rate is about right when the words fall into few topics and too fast when they fall into many. At $K = 32$ the floor stays on one grid width over a sixteenfold range of N, moves one step at the largest, and sits below the topic floor at four of the six N. Every synthetic retrieval floor lies below $8.33 \ln N$, so there the rule is a sufficient width. Figure 2 shows one such instance

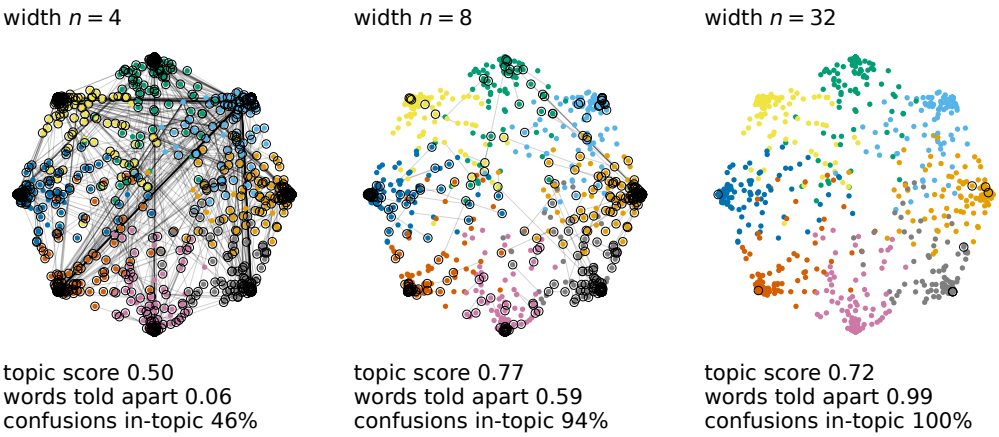

Figure 2: One synthetic vocabulary ($K = 8$ topics, $N = 1{,}000$ words, the middle of the N grid, first seed) embedded at three widths, in one fixed view: each word sits by its true topic mixture and is coloured by its dominant topic. A ring marks a word whose rank-n embedding is closer to another word's profile than to its own, and a line joins it to that word. At $n = 4$ confusions cross topics; at $n = 8$, the topic floor, 94% of them stay within a topic, so the embedding separates the topics but not the words; at $n = 32$, the retrieval floor, nine words remain confused. The rule asks for $8.33 \ln N = 58$. The embedding is rebuilt from the recorded generator settings and reproduces the recorded scores exactly.

Table 1: Slope of the floor against $\ln N$ (t-intervals), grouped by what the task must distinguish. The K=32 rows coincide: each floor sequence is one step, at opposite ends of the $\ln N$ grid.

	setting	b	95% t-CI	R^2	n
identity	synthetic words, retrieval (K=8)	$+7.25$	$[+3.98, +10.53]$	0.90	6
	synthetic words, retrieval (K=16)	$+5.28$	$[+0.32, +10.23]$	0.69	6
	synthetic words, retrieval (K=32)	$+1.85$	$[-1.12, +4.83]$	0.43	6
	`text8` words, retrieval	$+102.9$	$[+19.6, +186.1]$	0.84	5
	attention keys, recall	$+2.43$	$[+1.88, +2.98]$	0.93	10
category	synthetic topics (K=8)	0 (all 6 floors equal 8)		–	6
	synthetic topics (K=16)	0 (all 6 floors equal 16)		–	6
	synthetic topics (K=32)	$+1.85$	$[-1.12, +4.83]$	–	6
	node classification	no common growth (§3.3)			

word by word. What separates the red and blue synthetic series of Figure 1 ($K = 8$) is not N, the data or the embedding, but what the task has to tell apart.

3.2 Identity tasks grow with N, at rates the data set

On `text8`, the floor exceeds $8.33 \ln N$ at every N we ran (6/6), and the last failing grid width lies above the rule at 4/6 values of N at 95%, 5/6 at 98% and 6/6 at 99%: on real text the rule is not a sufficient width. Its requirement grows at $+102.9$ $[+19.6, +186.1]$ per unit of $\ln N$ (Table 1). The interval rests on five uncensored points; leave-one-out estimates range from +60 to +137, and 0/5 of the four-point fits exclude 8.33, so the separation from the rule is by magnitude rather than by interval. In the attention head, the 95%-of-peak threshold grows at $+2.43$ $[+1.88, +2.98]$ per unit of $\ln N$, and the smallest width reaching 0.999 at $+3.99$ $[+2.35, +5.62]$; both are well below the rule's rate. At $N = 4096$ the rule asks for 69.3 dimensions of a model whose embeddings have 64 in total and which solves the task to 0.999 (Table 2). The two real identity tasks, an SVD rank and a trained QK width, differ by a factor of 42; the rule's constant fits the synthetic one with few topics and neither real one. The attention slope survives eight times more training at 4 of 5 values of N, a ten-seed replication and every leave-one-N-out fit (+2.23 to +2.67); neither identity task favours the logarithm over a power law by AICc (Appendix B).

Table 2: Attention: the width at which recall is solved, against $8.33 \ln N$; $d_{\text{model}} = 64$ at every N.

N	8	16	32	64	128	256	512	1024	2048	4096
$8.33 \ln N$ (bold: exceeds d_{model})	17.3	23.1	28.9	34.6	40.4	46.2	52.0	57.7	63.5	**69.3**
smallest width with acc ≥ 0.999	2	2	3	4	6	8	11	16	16	32
95%-of-peak threshold	2	2	3	4	6	8	8	11	16	16

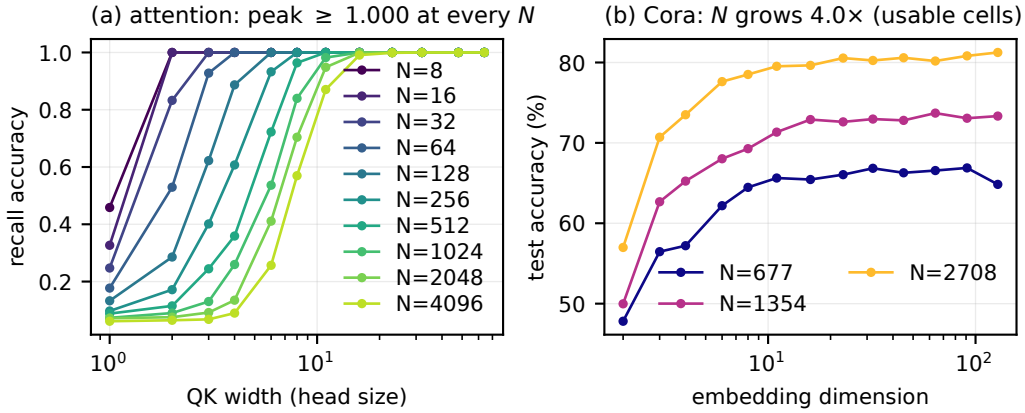

Figure 3: The curves the requirement is read from. (a) Attention shows a sharp cliff and a peak of 1.000 at every N, so the threshold is not measured against a depressed ceiling. (b) Cora, usable cells only: the collapse point stays between 6 and 11 dimensions at a 95% tolerance as N grows fourfold.

3.3 Category tasks on real graphs: no common growth

On the 12 usable cells (4 dropped for a test split under 200 nodes), the fixed-reference gap slopes are in Table 3. A requirement of $8.33 \ln N$ would make these gaps large — the task near chance at $d = 8$ — and would make g_{64} step up where N crosses 2,200. Instead the gap at $d = 8$ is a few points (Figure 3b) and no column moves it consistently with N. We quote the cluster-robust intervals as a width indicator rather than a calibrated test, since three clusters cannot support one. The three graphs do not share a slope: at $d = 8$ the within-dataset slopes are Cora $+1.71$, CiteSeer -2.14 and PubMed -0.11. The significance of the small pooled slopes depends on the reference dimension (at $d = 16$ the slope is negative for every reference, $p = 0.084$ at $d_{\text{ref}} = 128$, $p = 0.009$ at 64 and 91), and their sign on the sampler (Appendix B). Adjusting for $\ln n_{\text{train}}$ gives -0.47 at $d = 8$.

The floors themselves are small. At a 95% tolerance they are 2–32 dimensions, at or near the class-count bound $C - 1$ on PubMed and Cora (a description, with three graphs; Appendix C), against $8.33 \ln N$ of 49.9 to 82; the largest, 32, comes from the smallest and noisiest cell (CiteSeer at $N = 399$, seed-bootstrap range 16–64). At 99% the floors rise and the graphs diverge: Cora's grows by $+10.8$ per unit of $\ln N$, of the rule's order, CiteSeer's falls (-21.6) and PubMed's is flat ($+1.1$; Table 4). The width these tasks need depends on the graph and its labels, not on a common function of N, as Gu et al. (2021) find across networks; the category–identity contrast and the `text8` result do not move with the tolerance.

Entropy-guided selectors. The graph selectors built on the rule (Luo et al., 2021; Yang et al., 2023) prescribe roughly 100 to 130 dimensions for these graphs (98–123 in Luo et al. 2021, Table 4, $\lambda = 1$). On these benchmarks that is a safe choice. Across $d \in [48, 130]$ the spread of mean test accuracy is 0.17–1.25 points, comparable to the within-dimension seed standard deviation of 0.40–1.20 points. On Cora, 64 against 98 dimensions differs by $+0.03$, $+0.06$, $+0.16$ and -0.40 points across four models, so a value check cannot separate competing prescriptions here.

Table 3: Node classification: slope of $g_d = \mathrm{acc}(128) - \mathrm{acc}(d)$ against $\ln N$, in accuracy points (random induced subgraphs, GCN). t-intervals; cluster-robust column on t_2.

d	pooled OLS	dataset fixed effects	cluster-robust by dataset
4	-2.35 $[-4.62, -0.08]$ ($p = 0.044$)	-0.48 $[-1.29, +0.33]$ ($p = 0.212$)	-2.35 $[-5.96, +1.27]$ ($p = 0.108$)
8	-0.80 $[-1.89, +0.28]$ ($p = 0.130$)	-0.41 $[-1.74, +0.92]$ ($p = 0.499$)	-0.80 $[-3.05, +1.45]$ ($p = 0.265$)
16	-0.66 $[-1.43, +0.11]$ ($p = 0.084$)	-0.70 $[-1.68, +0.27]$ ($p = 0.135$)	-0.66 $[-2.64, +1.32]$ ($p = 0.288$)
32	$+0.01$ $[-0.55, +0.57]$ ($p = 0.968$)	-0.11 $[-0.85, +0.63]$ ($p = 0.737$)	$+0.01$ $[-1.07, +1.10]$ ($p = 0.970$)
64	$+0.11$ $[-0.31, +0.53]$ ($p = 0.566$)	$+0.06$ $[-0.49, +0.61]$ ($p = 0.813$)	$+0.11$ $[-0.86, +1.08]$ ($p = 0.671$)

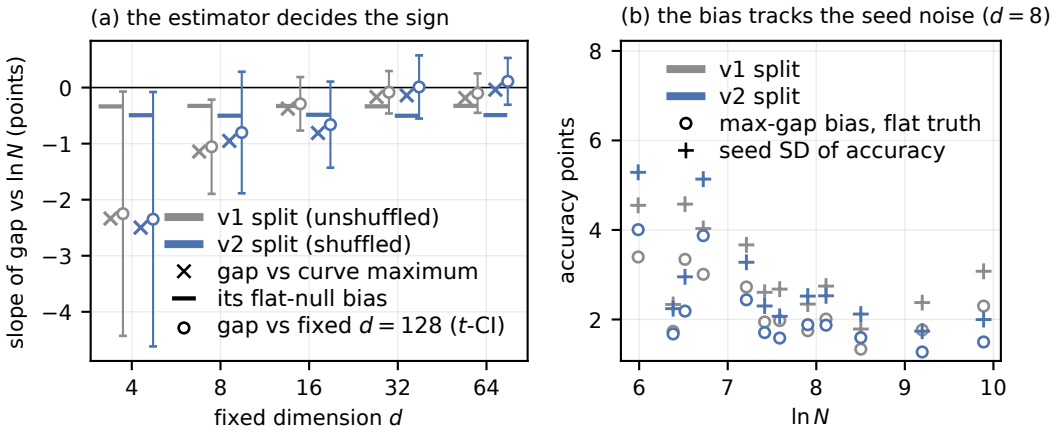

Figure 4: (a) Gap slopes on node classification under two estimators and two split routines (v1: unshuffled remainder; v2: shuffled). Crosses: gap against the curve's own maximum; dashes: the slope that estimator returns on a simulated *flat* truth with each cell's own noise; circles: gap against the fixed reference $d = 128$, with t-intervals. (b) The bias of the max-based gap under a flat truth, per cell, against $\ln N$, beside the seed standard deviation it tracks.

3.4 What the estimator does to the slope

The most natural pipeline for this measurement — a per-class split cut from one permutation, a gap measured against the curve's own maximum, pooled OLS — returns graph gap slopes that are *negative* at every fixed dimension, -1.14 $[-1.91, -0.37]$ at $d = 8$ ($p = 0.008$). The number is an artefact of two biases, both collinear with $\ln N$. The mechanism is the selection bias of a maximum over noisy estimates (Cawley & Talbot, 2010), which to our knowledge has not been reported for dimension-floor estimators.

Two biases. First, the maximum of 13 noisy dimension means exceeds the true plateau by an amount proportional to the seed standard deviation (one cell in Figure 5b, Appendix B). In this arm the seed standard deviation falls with N (-0.44 points per unit of $\ln N$, $p = 0.056$; -0.66 in the shuffled-split arm; Figure 4b), only partly because the test split grows. A flat truth therefore yields a negative slope: -0.33 per $\ln N$ at $d = 8$ on those cells, an upper bound, since the simulation draws the 13 means independently while real seeds share a split across d. On the records themselves, replacing the maximum by $\mathrm{acc}(128)$ moves the slope from -1.14 $[-1.91, -0.37]$ to -1.06 $[-1.90, -0.22]$, and the realised excess of the maximum over $\mathrm{acc}(128)$ has median 0.29 points over the 12 cells (largest 2.05). The selection bias is real, then, but accounts for a small part of the slope; the rest comes from the split, the pooling and the inference. Second, a split routine that cuts validation and test from the *unshuffled* tail of the permutation used to fill the per-class training quota leaves the validation set enriched in the classes that filled their quota first, and the test set in those that filled last. The total-variation distance between the two class distributions, median over seeds, reaches 0.26 on the usable cells and 0.39 overall, against 0.016–0.134 once shuffled, and model selection then happens on the distorted set. That distortion also shrinks with N. Neither bias is visible at a single N.

Removing them. On the same records, adding dataset fixed effects to the fixed-reference gap gives -0.74 $[-1.64, +0.15]$ ($p = 0.090$), and cluster-robust inference gives $p = 0.128$. With the shuffled split and the fixed-reference estimator (Table 3), the flat-null bias at $d = 8$ is -0.50 per $\ln N$ for the max-based gap and -0.02 for the fixed-reference gap, and the measured slopes are -0.95 $[-1.90, +0.00]$ and -0.80 $[-1.89, +0.28]$. The slope's sign is then no longer a property of the estimator; what remains is between-graph heterogeneity. Figure 4a shows all four combinations.

The general point. Maximum-relative floors are the default estimator. The remedies are cheap: a fixed reference, not a selected maximum, a flat-null simulation beside every slope, t-intervals, clusters treated as clusters, and one analysis path for text and figures.

4 Discussion

What this means for choosing n. For a task that separates K categories, the results point to sizing the width from what must be separated rather than from N: K on synthetic topics with $K \leq 16$, and 2–32 dimensions on the three graphs whatever the subgraph size. For a task that tells N items apart, the requirement grows with $\ln N$ at a rate the data set ($+2.43$ to $+102.9$ on the two real identity tasks), so the constant has to be measured, which the slope test does on subsamples of the data at hand. A rule derived for one task can still be a safe width for another (§3.3).

Criteria for choosing n. Yin & Shen (2018) derive a bias–variance criterion through the PIP loss; Hung & Yamanishi (2021) use sequential normalized maximum likelihood; Luo et al. (2021) and Yang et al. (2023) use graph entropy; Su (2020) gives the entropy argument of §2. They differ in machinery but are validated the same way: fix one corpus or a handful of graphs, compute the criterion, and check that its output lands near where a grid search lands at that N. Several contain N analytically and so make a slope prediction, and published tables already contain two such trends. Hung & Yamanishi (2021) select 65 dimensions on `text8` and 130 on Wikipedia, a doubling from the smaller corpus to the larger; Luo et al. (2021) select $98/101/123$ across graphs of 2.7k to 19.7k nodes, a rise that follows the $\ln N^2$ term of the rule. Our identity result is consistent with the first; our category result suggests that for classification the N-dependent term is not what sets the width. Re-running a field's validation under one controlled protocol has changed conclusions before, for graph classifiers (Errica et al., 2020) and metric learning (Musgrave et al., 2020); the slope test is such a protocol for dimension rules.

Capacity bounds. JL (Johnson & Lindenstrauss, 1984; Dasgupta & Gupta, 2003) gives $O(\ln N)$ for distance preservation and is tight (Larsen & Nelson, 2017); Seshadhri et al. (2020) argue near-linear rank is needed for triangle-rich networks; Bhojanapalli et al. (2020) show a low-rank bottleneck in multi-head attention when the head size falls below the sequence length; Likhosherstov et al. (2023) show that a width of $O(\ln L)$ lets self-attention approximate sparse patterns over L positions; and Sanford et al. (2023) show that a width growing as $q \ln N$ suffices for sparse retrieval at fixed precision, and one independent of N at infinite precision, which our width-2 construction instantiates (§4). These bound what can be *represented*, not what a task *needs*, and they differ from each other by orders of magnitude; a slope study works in the gap between the two. For categories, a random projection preserves every K-means cost to within $1 + \varepsilon$ with a dimension that depends on K and ε but not on the number of points (Makarychev et al., 2022). Cabannes et al. (2024) derive scaling laws in sample size and parameter size for associative memories built from outer products of embeddings; our attention task is retrieval of this kind, measured at the width where optimisation finds a solution. Pope et al. (2021) find that image datasets of lower intrinsic dimension, which lies far below the pixel count, are easier to learn; ours is the analogue for width, with the task in the role of the data.

The closest prior work. Two lines treat N-dependence as the object of enquiry. For retrieval, Reimers & Gurevych (2021) show that dense retrieval degrades faster with index size at lower dimension, and Weller et al. (2026) find that the number of documents a free embedding of width d can serve for all top-2 queries grows as a cubic polynomial in d, so the width needed grows as a power of the number of documents; our identity results agree in direction, and our functional-form test does not favour the logarithm either (§3.2). For graphs, Gu et al. (2021), across 135 real networks from $N = 16$ to 1.97M plus controlled synthetics, they find that for sufficiently large networks the selected dimension shows *no clear dependence on* N, tracking density and structure instead. **Our**

node-classification result confirms theirs in a narrower, more tightly controlled setting. We differ in holding the task and the supervision fixed while subsampling a single graph, in testing a named rule's slope, and in contrasting category tasks with identity tasks on the same embeddings.

Limitations. Five qualifications bound what these results show.

The attention axis is an optimisation threshold. A QK width of 2 suffices in principle: with the keys at distinct angles on a circle in two of the 64 coordinates and a scaled identity query, the softmax selects the matching key for any N once the scale is large enough, the infinite-precision case of Sanford et al. (2023). The representational floor is therefore a constant, and the attention floors in Figure 1 and Table 2 measure where Adam finds a solution. That threshold rises with N and is largely not a training-budget artefact ($8\times$ more steps leaves it unchanged at 4 of 5 values of N; Appendix B), but it is a statement about optimisation, not capacity. The comparison with d_{model} at $N = 4096$ is unaffected, because it bounds what the model *did* represent.

Context length and vocabulary are welded. Each position holds a distinct key, so N is both the sequence length and the key vocabulary. At length 512, raising the vocabulary to 4096 moves the threshold from 8 to 11 (2 seeds), so part of the slope is vocabulary. Narrowing the 64-wide value path to the head width gives 11, and reducing the value vocabulary from 16 to 4 gives 11, so the value path also affects where Adam finds a solution (Appendix B).

Confounds recorded, not removed. On random induced subgraphs mean degree rises with N (0.15 to 4.50 on the usable cells, slope +0.67) and the smallest cells sit near the percolation threshold. A breadth-first sampler holds degree within a factor of 2 with no trend (slope +0.01, $p = 0.859$) and gives gap slopes of the same order, with no positive significant slope in either arm; it changes the class mix, so we treat it as a second design (Appendix B). The labelled fraction cannot enter beside $\ln N$, being exactly $\ln n_{\text{train}} - \ln N$; the $\ln n_{\text{train}}$-adjusted slope is in §3.3.

Tolerance and small n. Floors are repeated at 98 and 99% (Appendix B); the category–identity contrast and the `text8` result hold at all three, while the graph floors rise and Cora's then grows with N. The `text8` slope rests on five uncensored points, the attention slope on ten and the graph regressions on three graphs, so no interval should be read to two significant figures, and 2.7 decades of N cannot cleanly separate $\ln N$ from $N^{0.3}$.

Scope and novelty. The synthetic generator's co-occurrence has rank K, so a topic floor at K is partly built in, and the synthetic results rest on one generator setting; the `text8` factorisation is only a relative of the model of (1). Our attention task, retrieval with learned keys, is the easiest case a head faces, so its constant is a lower bound for a production model. The graph finding confirms Gu et al. (2021). What we believe is new is the slope test for named rules, the same-embedding contrast of identity and category tasks, the attention measurement and the estimator bias (§3.4).

5 Conclusion

In this work, we tested dimension rules of the form $n > c \ln N$ on the claim they make: that the dimension a task needs grows by c per unit of $\ln N$. We held the task fixed and varied N in four settings, with an estimator whose expectation is zero when the truth is flat. The claim is compatible with the task the rule's derivation describes, telling items apart on clean data that fall into few topics, and does not transfer. The same kind of task grows more slowly when the items fall into many topics, and at rates more than an order of magnitude apart on real data, and tasks that only separate K categories need K dimensions on synthetic data with few categories and show no common trend with N on real graphs. The dimension an embedding needs is a property of what it must distinguish. Along the way, we found that floors read off a curve's own maximum acquire a slope in N that the data do not have, and that a fixed reference removes it. We hope this protocol lets future dimension rules be stated for a task and tested on their slope.

Reproducibility statement

Every statistic from our own sweeps (§3, Appendices A–B) is a macro generated from one analysis script's output, and the figures read the same file and compute nothing (the case study of Figure 2 rebuilds one recorded instance and checks it against its records; Figure 5 reads the recorded runs

directly); the numbers on entropy-guided selectors in §3.3 come from our separate reproduction of the MinGE and MGEDE tables, released with the same records. Appendix A lists grids, seeds and run counts; Appendix B the controls and tolerance checks; Appendix C the corrections four rounds of adversarial re-analysis forced on earlier drafts. The supplementary material contains the code for every run, the analysis and figure scripts, and the analysis output every number is generated from; the per-run records will be released with the code.

AI USE STATEMENT

This work was carried out by the authors in close collaboration with a generative AI assistant. Using the categories of the ICLR 2027 policy, among the uses whose disclosure is required we used generative AI to generate synthetic data sets (the associative-recall task and the planted-topic word-vector generator); to help develop theoretical models or conceptual frameworks (the readings of the rule in §2 and the packing reading in Appendix B); to formulate mathematical claims, limited to the informal arguments of the paper (the width-2 construction of §4 and the packing reading), since the paper states no theorems; to propose or refine hypotheses, including a two-requirements hypothesis that Appendix C records as withdrawn; to design or provide feedback on research methodology or experiments (the slope test of §2.1, the fixed-reference estimator and its flat-null control); to implement methods; to clean and reformat datasets (subsampling the citation graphs and `text8`); and to interpret results. We did not use it to provide critical ingredients for proving mathematical claims, to assist in the writing of proofs, to assist with translation, or to support qualitative and thematic data analysis. Among the uses whose disclosure is recommended, we used it to create or modify scientific figures, suggest experimental parameters, create or edit software code, create artifacts (the supplementary code and analysis output), draft parts of the paper, summarize and analyse existing literature, discover research topics and identify gaps, brainstorm, source and search for information, edit the paper for readability, identify relevant literature, format references, suggest the paper's structure, and propose its title and keywords. We did not use it to formulate questions for surveys or interviews or to transcribe recordings.

We have reviewed all AI-assisted work, and the paper's methodological finding came out of that review. Four rounds of adversarial re-analysis were run over the code, the recorded results and the draft, each finding verified independently before it was accepted; Appendix C lists what they changed and the claims we withdrew. The node-classification sweep was re-run from corrected code rather than re-analysed, the attention result was replicated at ten seeds with an independent evaluation seed, and every statistic from our sweeps is generated from recorded per-run outputs by a single script rather than transcribed. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International Conference on Machine Learning (ICML)*, 2020.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations (ICLR)*, 2024.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL `https://jmlr.org/papers/v11/cawley10a.html`.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. doi: 10.1002/rsa.10073.

Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. In *International Conference on Learning Representations (ICLR)*, 2020.

Weiwei Gu, Aditya Tandon, Yong-Yeol Ahn, and Filippo Radicchi. Principled approach to the selection of the embedding dimension of networks. *Nature Communications*, 12:3772, 2021. doi: 10.1038/s41467-021-23795-5.

Pham Thuc Hung and Kenji Yamanishi. Word2vec skip-gram dimensionality selection via sequential normalized maximum likelihood. *Entropy*, 23(8):997, 2021. doi: 10.3390/e23080997.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson–Lindenstrauss lemma. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 633–638, 2017. doi: 10.1109/FOCS.2017.64.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive flexibility of self-attention matrices. In *AAAI Conference on Artificial Intelligence*, pp. 8773–8781, 2023. doi: 10.1609/aaai.v37i7.26055.

Gongxu Luo, Jianxin Li, Hao Peng, Carl Yang, Lichao Sun, Philip S. Yu, and Lifang He. Graph entropy guided node embedding dimension selection for graph neural networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2767–2774, 2021. doi: 10.24963/ijcai.2021/381.

Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Performance of Johnson–Lindenstrauss transform for k-means and k-medians clustering. *SIAM Journal on Computing*, 2022. doi: 10.1137/20M1330701. Special section STOC 2019.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Nils Reimers and Iryna Gurevych. The curse of dense low-dimensional information retrieval for large index sizes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 605–611, 2021. doi: 10.18653/v1/2021.acl-short.77.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-1596.

C. Seshadhri, Aneesh Sharma, Andrew Stolman, and Ashish Goel. The impossibility of low-rank representations for triangle-rich complex networks. *Proceedings of the National Academy of Sciences*, 117(11):5631–5637, 2020. doi: 10.1073/pnas.1911030117.

Jianlin Su. Minimum entropy principle (VI): how should the dimension of word vectors be chosen? `https://kexue.fm/archives/7695`, 2020.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhenyu Yang, Ge Zhang, Jia Wu, Jian Yang, Quan Z. Sheng, Hao Peng, Angsheng Li, Shan Xue, and Jianlin Su. Minimum entropy principle guided graph neural networks. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2023. doi: 10.1145/3539597.3570467.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning (ICML)*, 2016.

Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

A REPRODUCIBILITY DETAILS

Run counts (recorded runs read by the analysis). Node classification, shuffled split, both samplers: 2,080. Node classification, first-draft split (used only in §3.4; unique runs): 1,040. Attention main grid: 360; step-budget control: 148; large-N replication: 160; auxiliary controls: 53. `text8`: 202 rank slices of 6 deterministic factorisations, each scored on two evaluation subsamples. Synthetic word vectors: 801 rank slices. Gradient-trained models in total: 3,841; word-vector rank slices: 1,003.

Node classification. Cora, CiteSeer and PubMed; GCN at every width d of the grid; subgraph fractions $f \in \{0.06, 0.12, 0.25, 0.5, 1\}$ (Cora, CiteSeer) and $\{0.01, 0.03, 0.1, 0.25, 0.5, 1\}$ (PubMed); random and breadth-first samplers; seeds 0–4. The sampler and the split draw from independent RNG streams seeded by the run seed. GCN: two layers, hidden width d, dropout 0.5, learning rate 0.01, weight decay 5×10^{-4}, 200 epochs, test accuracy read at the best validation epoch. Every row records N, split sizes, mean degree, per-class training counts, the validation/test class-distribution distance and the version of the split procedure.

Attention. 2,500 training steps, $N \in \{8, 16, \ldots, 4096\}$, $h \in \{1, 2, 3, 4, 6, 8, 11, 16, 23, 32, 48, 64\}$, seeds 0–2, Adam at 3×10^{-3}, batch 64, 16 values, evaluation on 2560 held-out queries drawn from the training seed. The step-budget control re-runs $N \in \{256, 512, 1024, 2048, 4096\}$, widths $\{3, 4, 6, 8, 11, 16, 23, 32\}$, seeds 0, 1, at 10,000, 20,000 steps. The large-N replication uses 10 seeds, a 19,968-query evaluation drawn from a seed independent of training, and widths $\{11, 16, 18, 20, 23, 26, 28, 32\}$ at $N \in \{2048, 4096\}$. The auxiliary controls use the same code at $N = 512$, widths $\{4, 6, 8, 11, 16, 23\}$, seeds 0–1: the published architecture, a 4096-key vocabulary, a value path as wide as the QK projection, and four values; plus seeds 3–7 at $N = 2048$, width 11, with the main grid's 2560-query evaluation.

Word vectors. `text8`: vocabulary = the N most frequent tokens, window ± 5, PPMI, truncated SVD to rank $n \in \{2, \ldots, 512\}$; retrieval = fraction of evaluated words whose rank-n reconstructed PPMI row is cosine-closest, among all N full-PPMI rows, to their own row (identity preservation). Synthetic: K planted topics, Zipfian frequencies, $N \in \{250, \ldots, 8000\}$, $n \in \{2, \ldots, 256\}$, 3 generator seeds; the topic score is the Spearman correlation between rank-n word similarity and the planted topic similarity. In full: word–topic mixtures are drawn from $\mathrm{Dirichlet}(0.3 \cdot \mathbf{1}_K)$, word frequencies follow a Zipf law with exponent 1 assigned to words at random, and topic weights p are drawn from $\mathrm{Dirichlet}(\mathbf{1}_K)$; the co-occurrence matrix is $C = AA^\top$ with $A = \mathrm{diag}(\mathrm{freq})\, W\, \mathrm{diag}(\sqrt{p})$, of rank K, and the embedding is the rank-n truncated SVD of its positive PMI. Retrieval is scored on up to 2,000 words and topic recovery on 20,000 random word pairs.

Entropy-guided selectors. The numbers in §3.3 come from a rerun of the MGEDE authors' released code, for its dimension estimate and for an accuracy sweep, on the Planetoid public splits of the three graphs: GCN, GAT at two training lengths and DAGNN, seeds 0–9, widths from 8 to 256 including the selectors' choices, eleven model–dataset curves in all. These numbers are transcribed from that reproduction's tables rather than generated by our analysis script.

Analysis. One analysis script reads every recorded run and writes a single results file; every number in this document is generated from that file, and the figures are drawn from it without further computation. Cells enter the graph analysis only with a complete 13-point grid, at least three common seeds and a test split of at least 200 nodes. Flat-null simulations use 4000 draws per cell.

B CONTROLS

Breadth-first sampler. Grows a connected region from a random node until the target size is reached. On the usable cells (12) mean degree spans 2.74 to 5.19 with no trend in N ($\ln$ degree $\sim \ln N$ slope +0.01, $p = 0.859$), against 0.15–4.50 (slope +0.67) under random sampling. The fixed-reference gap slope at $d = 8$ is -0.79 $[-1.70, +0.11]$ pooled ($p = 0.080$), -0.61 $[-1.81, +0.58]$ with dataset fixed effects, -0.79 $[-1.57, -0.01]$ cluster-robust ($p = 0.049$); the 95%-of-reference floors span 2–11. Across $d = 4/8/16/32/64$ the pooled slopes are -1.08 $[-2.76, +0.60]$ / -0.79 $[-1.70, +0.11]$

/ $+0.77$ $[-0.36, +1.90]$ / -0.65 $[-1.78, +0.49]$ / $+0.26$ $[-0.28, +0.80]$, cluster-robust -1.08 $[-1.65, -0.51]$ ($p = 0.015$) / -0.79 $[-1.57, -0.01]$ ($p = 0.049$) / $+0.77$ $[-3.26, +4.80]$ / -0.65 $[-2.63, +1.33]$ / $+0.26$ $[-1.27, +1.79]$: the two arms share the sign at some d and not others, the two cluster-robust intervals that exclude zero do so on the negative side (opposite to (2)), and nothing is positive and significant. The BFS region is not a uniform sample of the graph (its class mix differs from the full graph's) and it breaks the nominal supervision in the smallest cells: the smallest realised n_{train} is 41 and the smallest per-class count is 1 (random arm: 60 and 20). That is why we call it a second arm, not a degree-matched control. At the full graph the two samplers coincide, so the arms share their three largest cells.

Table 4: Every headline re-read at a stricter tolerance. Floors are the smallest grid width reaching the stated fraction of the reference accuracy. Graph slopes are descriptive (three to five points each). The attention cell uses the main grid's three seeds; the ten-seed replication reads 18 at 99%.

setting	quantity	95%	99%
synthetic topics	cells where the floor equals K (K=8/16/32)	6/6, 6/6, 5/6	6/6, 6/6, 6/6
`text8` retrieval	values of N needing more than $8.33 \ln N$	6/6	6/6
attention recall	floor at $N = 4096$ (rule: 69.3)	16	16
Cora	floor slope per unit of $\ln N$	$+0.0$	$+10.8$
CiteSeer	floor slope per unit of $\ln N$	-10.3	-21.6
PubMed	floor slope per unit of $\ln N$	-0.3	$+1.1$

How firm the attention slope is. Eight times more training leaves the threshold unchanged at 4 of the 5 values of N tested and lowers it by one grid step at $N = 256$ (8 to 6); at $N = 2048$ and width 11, mean accuracy on seeds 0, 1 is 0.9529 after 2,500 steps and 0.9502 after 20,000. A replication with 10 seeds, evaluated on 19,968 queries from an independent seed, returns the main grid's thresholds at the two largest N (16 and 16). Leave-one-N-out slopes span +2.23 to +2.67, and dropping the two points at the representational minimum gives $+2.85$ $[+2.15, +3.55]$; every variant excludes 8.33.

A packing reading of the two regimes. Telling N items apart at a fixed margin needs N nearly orthogonal codes, whose number grows exponentially with the width, so the width grows with $\ln N$ at a rate the data's margin sets. Separating K categories needs, in the linear case, the $K - 1$ directions spanned by the class means, however many items each class holds: the synthetic topic floors sit at K, and the graph floors at or near $C - 1$ on two of the three graphs.

Graph floors pooled. The pooled slope of the 95% graph floors against $\ln N$ is -5.53 $[-16.24, +5.17]$, and -5.53 $[-25.87, +14.81]$ with cluster-robust errors. That interval contains both 0 and 8.33: three graphs cannot pin a common slope down, which is why the text reports the gaps at fixed dimensions.

Tolerance. Every floor re-read at 98 and 99% of its reference. Planted topics: the floor equals K in 6/6 cells for $K = 8$, 6/6 for $K = 16$ and 6/6 for $K = 32$ at 99%. `text8` identity retrieval needs more than $8.33 \ln N$ at 6/6 values of N at 99%. The attention floor at $N = 4096$ is 16 at 95% and 16 at 99%. Graph floors at 99%, in increasing N: Cora 8, 16, 23; CiteSeer 64, 23, 16, 16; PubMed 4, 6, 11, 6, 8. Their slopes against $\ln N$ are $+0.0$, -10.3 and -0.3 at 95%, and $+10.8$, -21.6 and $+1.1$ at 99% (descriptive; three to five points each). At the largest N of each graph the rule exceeds the 99% floor by a factor of 10 on PubMed, 4 on CiteSeer and 3 on Cora.

Attention step budget. At $N = 2048$ and width 11, on the matched seeds {0, 1}, mean accuracy is 0.9529 at 2,500 steps, 0.9518 at 10,000 and 0.9502 at 20,000. The budget is not the binding constraint at $N = 2048$; at $N = 256$ longer training lowers the threshold from 8 to 6, and at the other values of N it leaves it unchanged. The 95%-of-peak threshold at $N = 2048$ is decided by width 11 at -0.0017 relative to the 0.95 bar (three seeds; bootstrap mode probability 0.70, exact over 27 resamples); it is the least certain point on the curve, which is why the large-N replication exists. Pooling the main grid's three seeds with five more (Appendix A) gives 8 seeds at width 11: mean 0.9471, s.d. 0.0070, s.e.m. 0.0025, one-sample t-test against 0.95: $p = 0.28$.

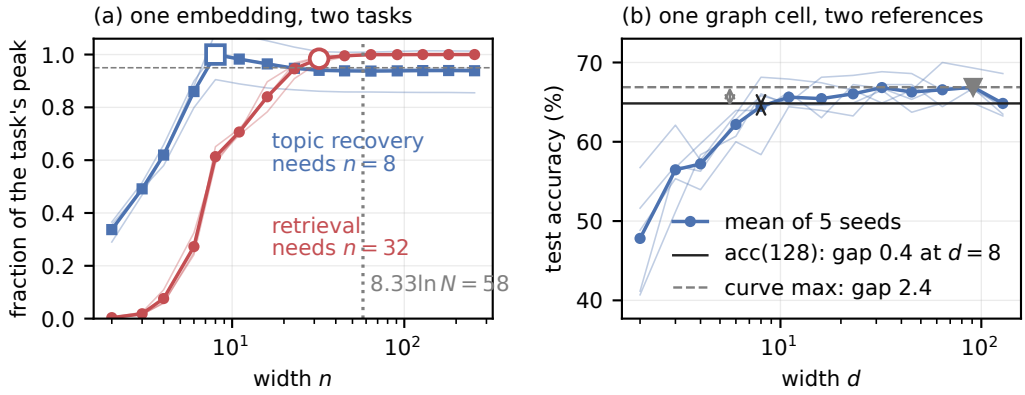

Figure 5: The same instance as curves, and one graph cell. (a) The set of synthetic word vectors ($K = 8$, $N = 1{,}000$, the middle of the N grid) scored on both tasks, each as a fraction of its own peak; thin lines are the three seeds, open markers the 95% floors. The same vectors need 8 dimensions to recover topics and 32 to tell words apart, both below $8.33 \ln N$. (b) The Cora cell whose seed standard deviation is the median of the usable Cora cells ($N = 677$); thin lines are the five seeds. The curve's maximum is a noisy high point at $d = 91$, so the gap at $d = 8$ measured against it is several times the gap against acc(128): the selection bias of §3.4 on one cell. The rule happens to select the cell with the largest realised bias of the 12.

Large-N replication. With 10 seeds, a 19,968-query evaluation from an independent seed and widths $\{11, 16, 18, 20, 23, 26, 28, 32\}$: at $N = 2048$ the 95%-of-peak threshold is 16 (main grid: 16; bootstrap 11: 0.09, 16: 0.91) and the smallest width reaching 0.999 is 18; at $N = 4096$ (10 seeds) the threshold is 16 (main grid: 16; bootstrap 16: 1.00) and the smallest width reaching 0.999 is 23. Mean accuracies ± s.e.m.: $N = 2048$: h11: 0.9465±0.0029; h16: 0.9984±0.0003; h18: 0.9995±0.0001; h20: 0.9995±0.0002; h23: 1.0000±0.0000; h26: 1.0000±0.0000; h28: 1.0000±0.0000; h32: 1.0000±0.0000; $N = 4096$: h11: 0.8699±0.0030; h16: 0.9896±0.0005; h18: 0.9942±0.0011; h20: 0.9979±0.0003; h23: 0.9994±0.0001; h26: 0.9999±0.0000; h28: 0.9997±0.0002; h32: 0.9999±0.0001. Su's requirement at these N is 63.5 and 69.3.

Attention estimator variants. Leave-one-N-out slopes span +2.23 to +2.67; excluding the two points at the representational minimum gives $+2.85$ $[+2.15, +3.55]$. Every variant excludes 8.33.

Functional form. Comparing $n^\star = a + b \ln N$ against $n^\star = cN^\alpha$ by AICc, both fitted by least squares in the space they are scored in, the preferred form is power on attention ($\alpha = 0.329$, ΔAICc $= 6.44$) and power on `text8` ($\alpha = 0.784$, ΔAICc $= 10.01$ on 5 points). On attention the verdict flips when a single floor moves by one grid step (1/20 such perturbations), and on `text8` it rests on five points.

Vocabulary versus length, and a real head. All runs here use $N = 512$, 2 seeds and 19,968 evaluation queries. We give the 95%-of-peak threshold and the smallest width reaching 0.999 for each: published architecture 8 / 11; key vocabulary raised from 512 to 4096 at fixed length 11 / none; value projection narrowed to the head width with the output read from it (a real single head) 11 / 16; value vocabulary reduced from 16 to 4 with the published architecture 11 / 16.

C Corrections forced by adversarial re-analysis

Four rounds of independent line-by-line audit were run on the code and the draft. We list what changed, so that a reader can see which numbers in earlier drafts should not be relied on:

- **Split construction.** Validation and test sets were cut from the unshuffled remainder of the per-class fill permutation, enriching validation in early-filling (majority) classes and test in

late-filling (minority) classes (§3.4). All node-classification records were regenerated with a shuffled remainder and an independent split RNG; the old records are used only in Figure 4.

- **Max-based gap.** The "estimator-free" gap used the curve's own maximum as the plateau; replaced by a fixed reference dimension, with a flat-null simulation reported.
- **Inference.** Intervals used 1.96 at $n = 4$–12; graph cells from one graph were treated as independent. Replaced by t-intervals, dataset fixed effects and CR1.
- **Form test.** The power model was fitted in log space and scored in raw space; AICc used $k = 2$. Refitted by nonlinear least squares with $k = 3$; undefined at $n = 4$.
- **Figures.** The figure code re-implemented the analysis without the test-size filter and with a different tolerance; three figure numbers in the earlier draft came from it. Figures now read the analysis JSON.
- **Withdrawn: "the structure floor declines with N"** — a sign error in the scoring function (below); and with it the two-requirements hypothesis built on it. The pair-sampling regime of that measure also determines its slope (a regime holding top-k fixed while N grows $32\times$ makes the scored pairs $32\times$ more selective), so no version of it is reported. A related artefact: a scoring function that forms $u = U\sqrt{S}$ from a truncated eigendecomposition of an indefinite PPMI matrix flips the sign of every negative eigendirection, so past the first negative direction each added dimension degrades the score and a "structure floor" declines with N.
- **Withdrawn: "the floor tracks the number of classes, $p = 0.013$"** — three datasets give an exact permutation p of 0.5. §3.3 notes only that the floors sit near $C - 1$ on two graphs, as a description.
- **Withdrawn: "3–11 dimensions"** — produced by an undefined estimator whose tolerance shrank with N; the defined estimator's range is reported in §3.3.
- **Withdrawn: "the attention floor at $N = 2048$ is undertrained"** — an intermediate correction of ours, refuted by an eight-seed replication (Appendix B); at matched seeds the step control leaves the threshold unchanged at 4 of 5 values of N and lowers it by one grid step at $N = 256$.
- **Text corrections.** Run count; 98–123 versus 99–130 attributed to their sources; the attention grid stated as 12 values; the direction of the degree confound (degree rises, not falls, with N under random subsampling).
- **Fourth round (text against records).** A sentence claimed the sign of the small graph slopes changes with the reference dimension; only their significance does. The synthetic topic floors are read against the curve's peak, not against $\mathrm{acc}(128)$ as §2.1 had implied; both readings are now reported. The $K = 32$ members of the synthetic series, one of which excludes 8.33, had been left out of the text and are now in Table 1. The training-budget control had been stated for every N; it holds at four of five. A seed-noise slope from the second split had been quoted for cells of the first. A 99% attention floor rested on three seeds; the ten-seed replication is now given beside it. The representational floor of the attention task, previously stated as a measured accuracy, is stated as the construction it is. The nearest retrieval work was added to the positioning.